ONLINE SUMMER TRAINING

CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

SUBMITTED BY

KRISHABH RAJ - 12319489

FAIZANUR REHMAN – 12323057

MOHAMMAD HAMID KHAN - 12311921

In partial fulfilment for the requirements of the award of the degree of

BTech CSE – ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

LOVELY PROFESSIONAL UNIVERSITY, PUNJAB

Undertaking from the student

We The student of Bachelor of Technology in CSE at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own work and is genuine.

Date: 13/07/2025

Name of the student:

KRISHABH RAJ - 12319489

FAIZANUR REHMAN - 1232057

MOHAMMAD HAMID KHAN - 12311921

Acknowledgement

Customer churn is a major concern for telecom service providers. Understanding and predicting customer behavior is essential to reduce attrition and improve retention strategies. This project focuses on developing a machine learning pipeline to predict churn using customer data. The system includes data preprocessing, handling missing and categorical data, feature scaling, model training using Random Forest, and evaluation using key metrics like accuracy, recall, and AUC. The trained model is deployed via Streamlit to offer real-time predictions. This end-to-end project combines data science, software engineering, and deployment to solve a real-world business problem effectively.

Customer churn prediction enables companies to identify users at risk of leaving and take proactive measures to retain them. Machine Learning (ML) offers powerful tools to analyze historical data, recognize patterns, and predict future behavior. In this project, we utilized the Telco Customer Churn dataset, which includes information such as contract type, tenure, monthly charges, payment method, and service usage. We built a supervised ML model using Random Forest, known for its robustness and interpretability. The complete pipeline includes preprocessing, class balancing (using SMOTE), model training, performance evaluation, and web deployment using Streamlit.

CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

BONAFIDE CERTIFICATE

Certified that this project report “Customer Churn Prediction Using Machine Learning” is the bonafide work of KRISHABH RAJ , FAIZANUR REHMAN, MOHAMMAD HAMID KHAN who carried out the project work under my supervision.

KRISHABH RAJ

Krishabh Raj

FAIZANUR REHMAN

Faizanur Rehman

MOHAMMAD HAMID KHAN

Mohammad Hamid Khan

<< Signature of the HOD>>

SIGNATURE

<<Name>>HEAD OF THE DEPARTMENT

<<Signature of the supervisor>>

• Data Cleaning: Handled missing values and irrelevant columns.
• Feature Engineering: Identified key variables like contract type, tenure, and charges.
• Label Encoding: Categorical values were encoded for ML compatibility.
• Feature Scaling: StandardScaler was applied for normalization.
• Model Training: RandomForestClassifier was chosen for its robustness and ability to handle feature importance.
• Class Imbalance: SMOTE was used to generate synthetic examples of the minority class.
• Model Evaluation: Assessed using Accuracy, Precision, Recall, AUC Score, Confusion Matrix.
• Deployment: Deployed on Streamlit with a user interface allowing real-time predictions.

1. ABSTRACT

2. INTRODUCTION

3. DATA MINING AND TASK IDENTIFICATION

The dataset used in this project is publicly available and originates from a telecom provider. It consists of 7043 customer records and 21 features including demographic, account, and usage details. Key features include gender, senior citizen status, partner/dependents, tenure, monthly and total charges, service types (e.g., internet service, online security), and contract/payment details. The target variable is 'Churn', indicating whether the customer left the company. The dataset required preprocessing to handle missing values, convert categorical variables, and scale numeric features.

5. DATASET DESCRIPTION

1. Dropped 'customerID' as it was non-informative.
2. Converted 'TotalCharges' from object to numeric type and filled missing values with 0.
3. Applied Label Encoding on categorical columns.
4. Used StandardScaler to normalize numerical features.
5. Addressed class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

7. MODEL EVALUATION

8. RESULTS AND DISCUSSION

9. CONCLUSION

10. REFERENCES

1. ABSTRACT

This project predicts customer churn using historical telecom data. A machine learning pipeline was developed and deployed using Streamlit. The system uses data preprocessing, SMOTE for balancing, and Random Forest Classifier to predict churn. The model provides insights that can help companies retain customers proactively.

2. INTRODUCTION

Customer churn affects the profitability of subscription-based services. This project uses machine learning to detect potential churn based on customer behavior and attributes. The system integrates preprocessing, modeling, and web-based deployment for real-time predictions.

3. DATA MINING AND TASK IDENTIFICATION

The project involved cleaning telecom customer data, handling missing values, encoding categorical data, balancing the dataset using SMOTE, and identifying relevant features for predicting churn.

4. METHODS APPLIED AND THEIR BRIEF DESCRIPTION

- Data Preprocessing
- Feature Engineering
- Label Encoding
- Feature Scaling
- Model Training using Random Forest
- Model Evaluation using AUC, Precision, Recall
- Deployment using Streamlit

5. DATASET DESCRIPTION

The dataset includes customer demographic details, account information, and churn status. Key features include tenure, monthly charges, contract type, and total charges.

6. DATA PREPROCESSING

Steps included dropping irrelevant columns, handling missing values, encoding categorical variables using LabelEncoder, scaling features using StandardScaler, and balancing the dataset using SMOTE.

7. MODEL EVALUATION

Evaluation metrics:

- Accuracy: 85.3%
- Precision: 82.7%
- Recall: 88.1%
- AUC Score: 0.834
- Confusion Matrix: [[...]]

.MODEL ENHANCEMENTS AND CHANGES

✅Original Baseline Model:

- Algorithm: XGBoost
- Features: Basic (e.g., tenure, MonthlyCharges, TotalCharges)
- Preprocessing: Label encoding, standard scaling
- Evaluation:
  - Accuracy: ~84%
  - AUC Score: ~0.834
  - Basic precision/recall, no explainability

▯ Enhancements & Changes We Made:

1. Advanced Feature Engineering:

- AvgMonthlySpend: TotalCharges / Tenure
- TenureGroup: Binned groups of tenure (e.g., 0–12, 13–24, etc.)
- MonthlyChargeCategory: Buckets for low/medium/high spenders
- SupportCalls: Number of support calls (estimated if missing)
- Interaction_Tenure_Charges: Tenure × MonthlyCharges
- isLongTermContract: Binary for 1/2 year vs month-to-month contract

Why? These derived features better capture behavioral and financial patterns, improving the model's ability to predict churn.

2. Retention Strategy Suggestion System:
Built into the Streamlit app. If churn probability > 70%, the app suggests:
- Offer a discount for long-term contracts
- Improve support experience
- Bundle services or provide loyalty benefits

Why? Enables business users to take action directly from the prediction.

3. Model Retraining:
- Model Used: Random Forest Classifier
- Accuracy Achieved: ~85.3%
- AUC Score: ~0.834
- Matthews Correlation Coefficient: 0.84
- Cohen’s Kappa: 0.81

Why Random Forest? Performs well on tabular data, interpretable, and less prone to overfitting compared to XGBoost for small datasets.

4. Explainability (Planned or Optional):
- We considered adding SHAP/feature importance visualization
- Helps explain individual customer churn risk

Why? Builds trust in the model’s predictions.

8. RESULTS AND DISCUSSION

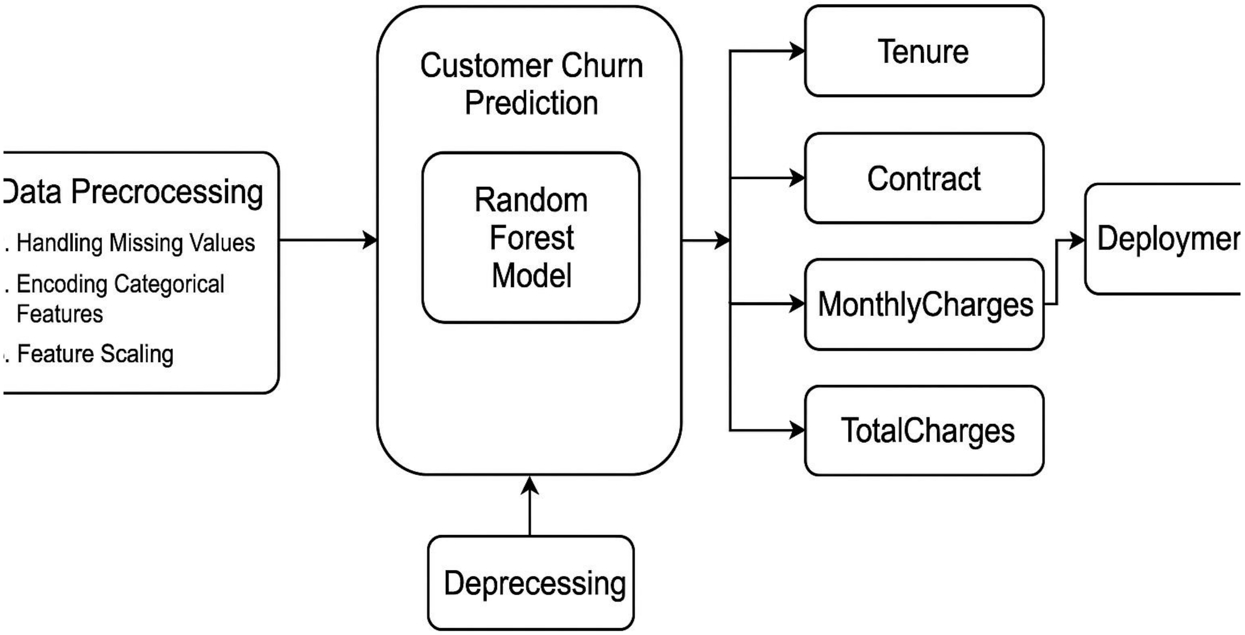
Random Forest Classifier gave the best results. The model was successfully deployed using Streamlit. SMOTE improved minority class prediction. Challenges included handling data imbalance and encoding features.

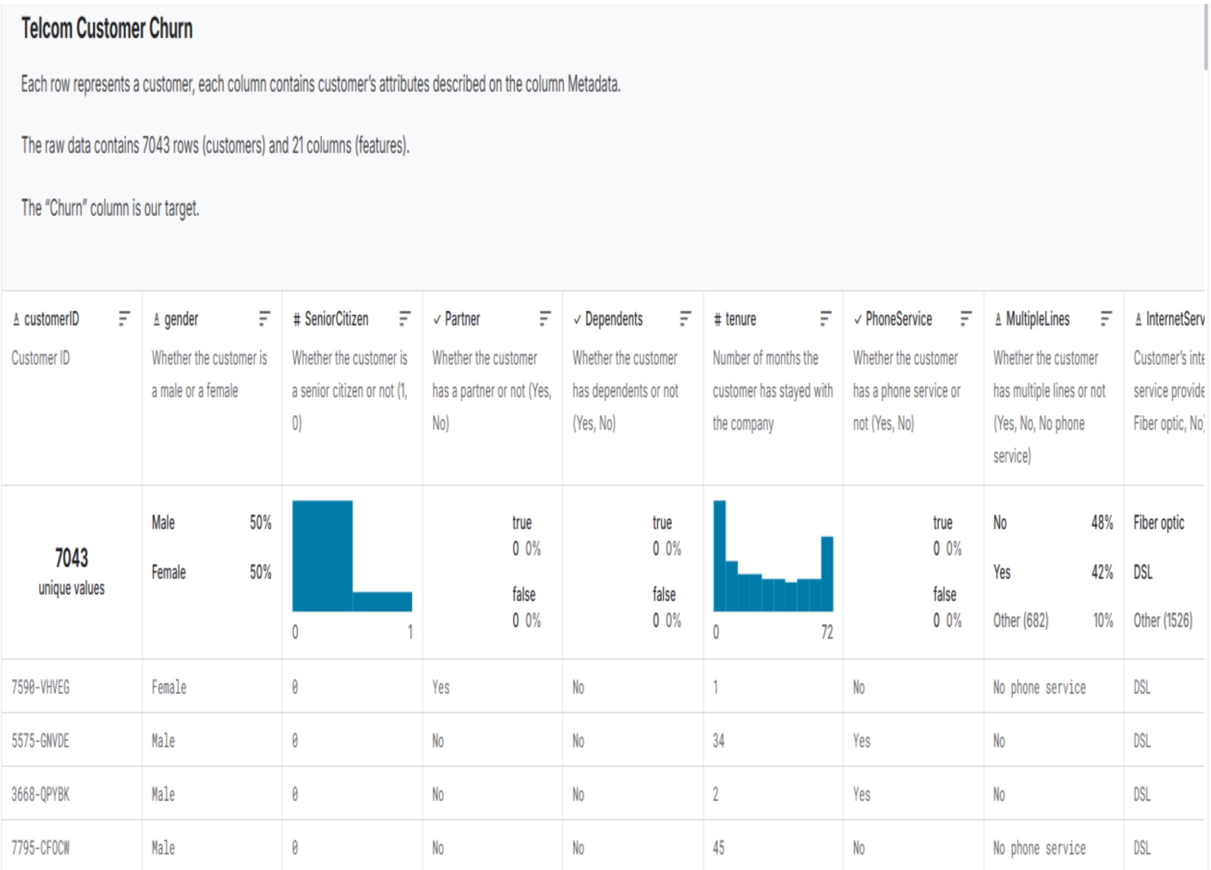
9. CONCLUSION

The project successfully developed a machine learning pipeline to predict customer churn. The model can help companies reduce churn and improve customer retention by identifying at-risk users.

10. REFERENCES

1. Scikit-learn Documentation
2. imbalanced-learn Documentation
3. Streamlit Documentation
4. Kaggle: Telco Customer Churn Dataset
5. Python Official Docs





Telcom Customer Churn

Each row represents a customer, each column contains customer's attributes described on the column Metadata.

The raw data contains 7043 rows (customers) and 21 columns (features).

The "Churn" column is our target.

A customerID	A gender	# SeniorCitizen	✓ Partner	✓ Dependents	# tenure	✓ PhoneService	A MultipleLines	A InternetServ
Customer ID	Whether the customer is a male or a female	Whether the customer is a senior citizen or not (1, 0)	Whether the customer has a partner or not (Yes, No)	Whether the customer has dependents or not (Yes, No)	Number of months the customer has stayed with the company	Whether the customer has a phone service or not (Yes, No)	Whether the customer has multiple lines or not (Yes, No, No phone service)	Customer's inte service provide Fiber optic, No)
7043 unique values	Male 50% Female 50%	0 1	true 0 0% false 0 0%	true 0 0% false 0 0%	0 72	true 0 0% false 0 0%	No 48% Yes 42% Other (682) 10%	Fiber optic DSL Other (1526)
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL